面向域间路由优化的覆盖网技术

李彦君　张国清

摘要：互联网的分布式管理导致了其整体传输质量无法保证的缺陷，域间路由优化覆盖网成为解决这一问题的热点技术。本文回顾了互联网路由现状，对现有覆盖网类型进行了细化归纳，分析了各类技术的优点与缺陷，在此基础上，探讨了域间路由优化覆盖网技术所面临的网络冲突、流量震荡等各种问题，并对覆盖网路由性能改善做了深入的量化综合分析，最后，总结了实现一个具有实用价值的高效、可扩展的域间路由优化覆盖网所面临的关键技术问题。

1　引言

以 TCP/IP[1]技术为核心的互联网的发展起源于军事上对抗打击的需求。它强调各个子网的独立性、分布性与自治性。历经 20 多年的商用化发展，互联网凭借其良好的统一性、兼容性在网络通信领域内确立了无可撼动的地位，其爆炸式的增长与应用是其设计者当初都始料未及的。

如今，互联网规模仍然在飞速扩张。虽然数据交换与路由等软硬件技术时刻都在创新与进步，但相对当初而言，互联网分布式、独立性的特点却一直没有改变。庞大的互联网分别为许多不同的机构与运营商控制与管理，它们之间的利益与差异已成为互联网演进的严重桎梏。如近年发生的台湾光缆断裂、巴基斯坦电信局封杀 YouTube 操作[2]等大量全球性故障事件就揭示了表面上看似日趋成熟完善的互联网仍极为脆弱的一面。

为此，互联网相关的机构和组织作了许多努力，试图克服互联网的脆弱性，并为各类业务提供相应的传输质量保证。早在上世纪 90 年代，IETF[3]组织就意识到了互联网这方面存在的缺陷，提出了 DiffServ、InterServ 等多种整体服务质量（quality of service，QoS）架构方案[1][2]。然而，这些方案的实现都需要满足两个基本条件：第一，沿通信链路上的所有路由器都要进行统一协调的业务调度和缓存控制；第二，端到端的网络流量穿越多个不同互联网服务提供商（Internet Service Provider，ISP）维护管理的网络域时，需要互联网服务提供商的积极配合与协调。而出于商业利益的考虑，互联网服务提供商只希望能以最小的投资获得最大的市场收益，每个互联网服务提供商均只愿意对本网络域内的传输质量和抗毁能力提供服务保证而不愿为提高整体服务质量增加负担。

因此，迄今为止，几乎没有任何一种有效的整体方案能够在现有互联网上付诸实施，而另一方面，大量新兴应用如音视频交互、交互式网络电视（IPTV）、网络游戏等实时性要求较强的业务已日趋成为网络主流内容，它们对 IP 网络的整体服务质量需求越来越强烈。如何为跨接多个网络域的业务传输（即域间传输）提供令用户满意的整体数据传输性能，已成为当前网络技术研究极富挑战性的焦点问题之一。

显然，在当前这种庞大的网络环境下，局限于单独一层的 IP 路由协议研究已经很难解决问题。新一代网络路由技术需要寻求一种中间层的模型机制，来提供跨越不同网络、不同运营商与不同管理域的端到端的业务能力，同时又保证对已有 IP 协议的前向兼容。在这样

1 Transmission Control Protocol/Internet Protocol，传输控制协议/互联网络协议
2 2006 年 2 月 22 日巴基斯坦政府在国内封锁 YouTube 网站的行动使全球大多数互联网用户 2 月 24 日有好几个小时不能访问 YouTube 网站
3 Internet Engineering Task Force，互联网工程任务组

的环境下，出现了架设在以 IP 层为基础之上的覆盖(Overlay)网络，并逐渐成为解决这些问题的主流方式。近 10 年来，国外研究机构先后提出了大量具体的网络结构与构建方案：如 Detour、弹性覆盖网(Resilient Overlay Network，RON）、QRON、SON、NIRA (New Inter-Domain Routing Architecture，新型跨域路由结构) 等等，但由于各种各样的原因制约，这些方案至今均未能实现大规模的应用部署，网络的发展再次在十字路口踟蹰不前。

本文试图针对这些域间路由优化技术进行分类探讨，总结其优缺点，在此基础上对覆盖网技术在域间路由优化方面的可行性与关键难点问题做深入的综述分析，本文其余部分组织结构如下：

第二部分回顾了现有路由优化覆盖网的研究现状，并对其类型进行了完全分类，小结了各类型覆盖网的技术优势与缺陷；在此基础上，第三部分讨论了面向域间路由优化的覆盖网技术所面临的各种问题与挑战；第四部分分析了实现一个高效、可扩展域间路由优化覆盖网的关键技术问题；第五部分是结论与展望。

2 覆盖网研究现状与分类

覆盖网络的概念最早可以追溯到早期的互联网。从某种角度讲，互联网就可以看作一个巨大的覆盖网(overlay network)，以 IP 层作为覆盖层，利用 IP 协议，跨越以太网、令牌网等异构的底层网络提供“尽力而为”的数据传输服务。覆盖网的兴起，是缘于现有互联网域间路由能力已越来越难以满足各类被承载业务的需求。

2.1 互联网域间路由现状

早在 1995 年，帕克森（V. Paxson）等人就开始对方兴未艾的 Internet 网域间路由性能进行了统计[3]，结果发现互联网上的网络大约存在 3.3%的严重路由故障率，大于 20%的路径故障无法在 10 分钟内得到修复，而拉伯维茨（C. Labovitz）等人在 1997 年至 2000 年的统计表明:现有互联网中 10%的路由器可靠性不到 95%，65%的路由器可靠性小于 99.9%，且故障恢复时间常达 15 分钟之久，40%以上的路由失效甚至需要 30 分钟以上才能恢复正常[4]。钱德拉（Chandra）[52]等人通过主动探测发现互联网络中 5%的故障甚至会持续 2.75 小时以上。

由于历史原因，现有的互联网实际是由许多个相对独立的自治系统（AS：Autonomous System）通过域间路由协议(BGP：Border Gateway Protocol)互联而成的一个庞大的网络。每个自治系统一般由一个独立的实体来控制。这些实体大部分是商业化组织，如互联网服务提供商。BGP 协议是当前互联网自治系统之间采用的一个边界网关路由协议，作为互联网最核心的构件，是所有自治系统之间传递网络信息的基本机制与互联纽带，同时也是互联网服务提供商实现策略控制的主要手段，对互联网的演化起至关重要的作用。但 BGP 协议自 1989 年产生至 1995 年修订为第四个版本后，迄今为止，几乎没有发生过大的改变。

出于扩展性考虑，BGP 协议只采用简单的跳段作为度量，这导致所谓的最佳路径并不是路由性能最优[5]；而为了避免频繁更新路径引起的“路由翻动”现象，BGP 不对路径性能进行实时探测，且采用单一的路径选择。这导致链路出现拥塞时，业务流量无法避开拥塞点，发生故障后，又呈现严重的慢收敛特性[6]。

这些测量结果与研究结论已充分说明：当前端到端的互联网域间路由机制既不保证端主机间的最佳传输质量，也不能迅速对故障进行快速恢复或响应。为了弥补 BGP 协议的缺陷，研究人员先后对 BGP 协议提出了多种扩展方案[8-10]。然而，这些设想和建议在实现上都需要进行大规模的额外升级部署，缺乏过渡性。

2.2 覆盖网研究现状及分类

2.2.1 覆盖网应用现状

为了克服互联网域间路由自身的缺陷，满足层出不穷的新兴业务对网络安全性、稳定性、可靠性等多方面的要求，针对具体的业务类型，人们构建了各式各样的专用覆盖网，如：用于发布存储流媒体数据的内容分发网络 (Content distribution network， CDN)网络[11]、用于应用层组播的终端系统组播技术 (end system multicast，ESM) [12] 与 Overcast 系统[13]、用于文件共享的对等传输（Peer to Peer，P2P）网络[14]以及仿真实验网络 Plantlab [15]和各种业务组合网络等[16]。

这些网络都是由一系列分布于互联网各个自治系统内部的服务节点以及连接它们的逻辑链路所构成的虚拟网络。通过服务节点的转发，用户终端能更有效地利用互联网资源，为相应业务提供现有互联网路由所无法满足的性能需求。这种覆盖网络实现方式的最大优点在于可以灵活方便地部署在基础网络之上，并不需要改变现有网络架构。

近年来，受这些应用覆盖网的研究启示，支持服务质量路由的覆盖网迅速成为解决互联网跨域传输故障收敛慢、稳定性差等问题的热点技术。路由覆盖网作为一种中间层模型机制，可通过节点间彼此协作主动选取高效路径路由数据包来提升端到端网络传输质量，满足用户应用性能需求，提供更为可靠、容错性更好的服务[17][18][19]。在弹性覆盖网和 Detour[20]等典型路由覆盖网技术中，研究者们已通过建设相应的实验网，验证了采用路由覆盖网在进行跨域传输时，相对 BGP 协议在快速响应、故障恢复、服务质量保证等方面的巨大优势。

2.2.2 域间路由优化覆盖网的分类

依据覆盖网实现技术的不同，我们可以将这些支持跨域服务质量路由的覆盖网络分为中继型覆盖网和虚拟网络（virtual network，VN）型覆盖网两个大类，而依据组网节点位置与系统实现层次，又可以进一步细分为网络层覆盖网与应用层覆盖网。

（1） 中继型路由优化覆盖网

中继型覆盖网的原理是发送方把数据包发给一个中继节点，中继节点再将数据包转发给接收端，从而形成一个多跳覆盖网路径。由于两跳的覆盖网路径提供的性能、可靠性及路径多样程度（path diversity）与多跳覆盖网路径提供的近似[21]，因此大多数覆盖网研究出于简化和易于实现的目的，均采用经过一个中继节点转发的两跳覆盖网路径。

■ 弹性覆盖网

弹性覆盖网是中继型覆盖网的典型代表 [17] 。按系统的实现层次，它是构筑在现有互联网路由层次之上的一个应用层覆盖网。在其上构建的应用程序可以利用建立在底层网络上的弹性覆盖网路由库，在现有的互联网上对路径重新做出选择。弹性覆盖网的体系结构如图 1 所示。它采用了两跳中继的应用层路由方式。其系统节点监测彼此之间的路径功能、性能，定期检查自己和其他节点之间的网络通路情况。应用层可通过对选取不同中转节点产生的多条转发路径的综合延迟、丢包率和吞吐量等多维权值进行评估后，选择可满足业务服务质量

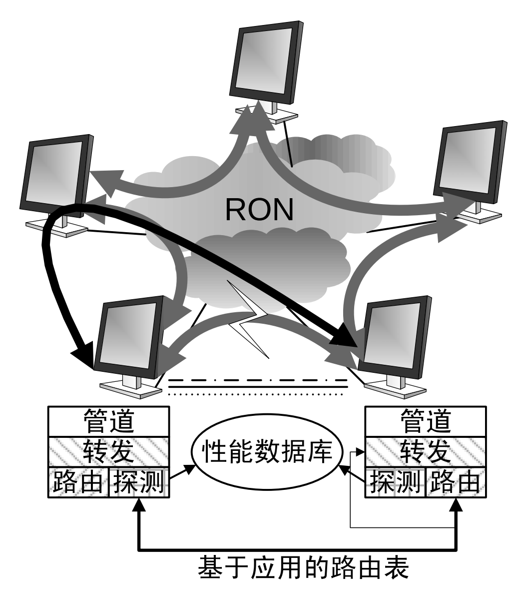


图1. 弹性覆盖网系统结构

需求的转发路径。同时，弹性覆盖网系统采用的应用层路由方式还可利用跨域传输时 BGP 协议所不能发现或无法利用的隐藏路径进行快速的路由故障恢复，提高包传输的可靠性。虽然弹性覆盖网系统在点对点之间传输优化的效果较好，在较小规模的网络环境下也可以有很快的反应，但是，弹性覆盖网也存在很多不足之处：

首先，弹性覆盖网网络的节点之间要以固定时间间隔以完全网状连接探测的方式大强度地发送探测包。这种小间隔、频繁的探测给网络带来巨大的开销，极大地限制了弹性覆盖网网络的可扩展性。

其次，弹性覆盖网的虚拟网络采用静态管理，转发节点限于少数服务器节点，结构不够灵活，无法充分有效利用隐藏路径和私有路径。

最后，弹性覆盖网的路径选取仅仅是依据当前洪泛探测的邻居节点中性能比较优异的节点中转，并没有参照底层拓扑，因此覆盖网路径与底层的默认 IP 路径的相关度就较高，以致对互联网上出现的某些性能下降和路径故障无能为力。

- QRON（QoS-Aware Routing in Overlay Networks）

为了解决弹性覆盖网系统的扩展性与针对具体业务的性能优化问题，李智（音译，Z.Li）等人提出了 QRON（QoS-Aware Routing in Overlay Networks）体系结构[22]，试图用一个通用的覆盖服务网（Overlay Service network，OSN）来满足各种应用层服务的需求。

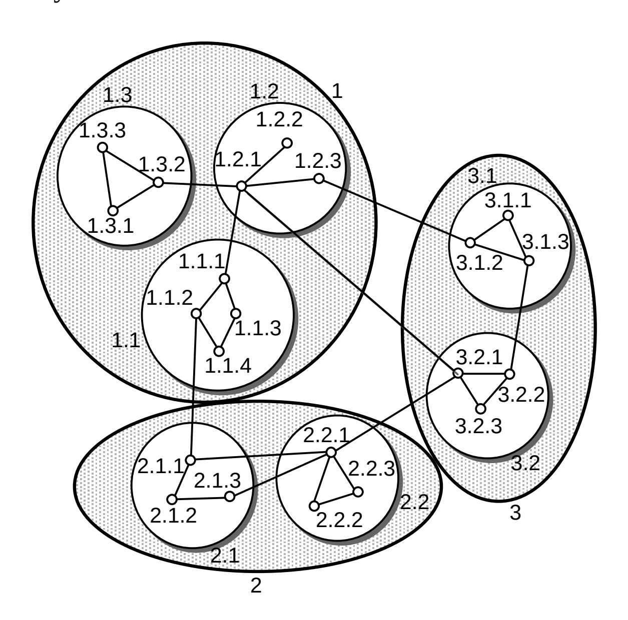


图2. QRON 体系架构

QRON 利用 OBs(Overlay Brokers，覆盖中介)来建立覆盖网服务网，这些 OBs 由覆盖网服务提供商部署在互联网的各个自治系统中，彼此协作形成覆盖服务网，为上层应用提供服务，比如资源分配和协商、覆盖网中继路由、拓扑发现等。QRON 专注于设计一个服务质量感知的路由协议，主要目的是为上层服务质量敏感的覆盖网应用找到满足服务质量要求的覆盖网路径，同时进行覆盖网流量均衡。由于 QRON 中的 OBs 采用的是层次化组织的方式，依据“网络距离”进行分簇，使得 QRON 有比较好的可扩展性。

相比弹性覆盖网，QRON 是一种带有虚拟网络性质的中继覆盖网，它采用的层次化部署可扩展性较好；提出的基于路径性能进行路由选择的算法使得选路效率较高；同时，它设计了通用的感知路由协议，具有较好的普适性。但是，QRON 是基于网络层的覆盖网，它的所有节点都需要自行部署，代价很高；此外，QRON 在进行路径性能评估时仅仅测量了丢包、延迟和可用带宽，并且测量方法准确度不高，这种不可准确测度的可用带宽设计的路由算法使其通用性受到了极大的限制。

- Spines 覆盖网络

Spines 是由约翰-霍普金斯大学的分布式网络实验室所开发的通用覆盖网络系统[25][26]，其源代码公开，可以用来进行覆盖网络协议的测试和开发。Spines 提供了两

级的层次结构。用户程序要连接到最近的覆盖节点。这个覆盖节点通过覆盖网络发送或转发数据到目的地址。Spines 运行于应用层，不需要改变基础设施和核心接入，因此与弹性覆盖网一样具有良好的可部署性。其分为两级的好处是可以限制覆盖网络的规模，以减少交换流控信息的数据量。覆盖节点既是服务器，用于连接各种各样的应用，同时也是转发路由器，用于节点之间传输数据包。由于 Spines 这样的等级结构，使得应用既可以位于覆盖节点处，也可以连接到与节点相连的其它机器上。显然，Spines 相比弹性覆盖网，在扩展性方面有了很大的进步。通过 Spines 系统，业务可实现低开销的可靠传输，可改善实时传输的时延、丢包率和吞吐量等特性，可以支持对服务质量有较高要求的多媒体业务，如网络电话（VoIP）视频会议等。

但 Spines 网络也有自身的缺点：一方面，Spines 只支持固定节点的拓扑结构，节点不能动态地加入和离开 Spines 网络。一旦某个节点或者链路失效，可能使得节点之间需要绕很远的路径来传输信息，也可能导致整个网络崩溃。另一方面，Spines 建立的拓扑结构往往不是最优的，甚至对底层网络的实际拓扑结构没有任何了解，传输灵活性和效率比较差。

（2） 虚拟网络型路由优化覆盖网

虚拟网络型覆盖网与中继型覆盖网一样是在网络关键部位设置节点。但虚拟网络型覆盖网中的节点与中继型节点不同的是其兼具监测底层网络资源分布、带宽利用、链路性能、流量负荷等功能，并可依据监测结果进行动态资源分配，从而实现网络性能的整体优化。虚拟网络型覆盖网可以理解为“虚拟的覆盖专用网”，终端用户流量可以在该虚拟网范围内得到传输优化。

- 绕路路由（Detour）

萨维奇（S. Savage）等于 1999 年提出的 Detour 路由系统，可以认为是早期的虚拟网络型路由覆盖网。它由一组分布式的边缘路由节点通过隧道技术连接而成，进入 Detour 入口节点的流量数据被重新封装为新的 IP 分组，沿互联网传输至 Detour 出口节点，最终传输至用户端（如图 3 示）。

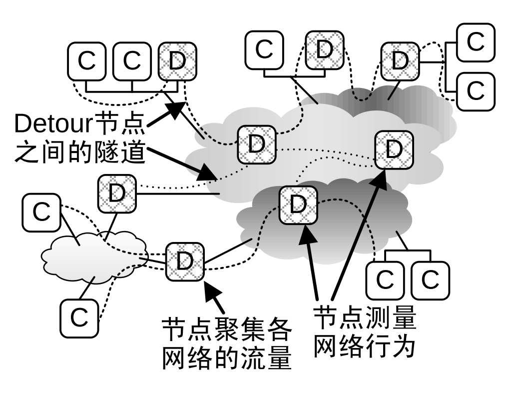


图3. Detour 路由框架

Detour 试图在大规模网络范围内研究性能敏感的新路由算法，以改善互联网的网络性能。它主要包括虚拟网络管理与节点路由控制两部分功能。用户节点通过将自身缺省路由器改为靠近用户的边缘 Detour 节点而进入 Detour 虚拟网，通过它将流量传输至目的节点。

相比互联网路由，Detour 系统在分钟级别的时间尺度上进行路由器选取调整，由于采用了迂回路由、单流多径传输与包分类等技术，可以在一定程度上改善网络端到端的传输性能。但是其开销与限制性也显而易见：首先，Detour 虚拟网部署于底层路由节点之上并运行独立的算法，要在互联网范围内形成可应用的统一大规模网络，部署代价将非常高昂；其次，即便缺省的互联网路由效率更高，使用 Detour 的用户流量也仍将全部经由 Detour 网络进行传输。这加重了 Detour 的负荷，使它面临和原互联网同样的扩展性与高效性矛盾的老问题；此外，Detour 的网络结构缺乏灵活性，无法充分有效地利用互联网的大量隐藏路径。

- OverQoS

如果说 Detour 网络还兼有中继型覆盖网络的特点，那么 OverQoS[23]就是一个非常典型的覆盖网虚拟覆盖网络。在 OverQoS 的架构下，服务供应商向传统的互联网服务提供商购买网络访问权，然后在不同的路由域中部署 OverQoS 节点组成一个覆盖网。在传统服务质量机制中，IP 路由器负责控制分组缓冲区和输出链路带宽，而 OverQoS 节点不考虑底层性能，它引入了可控丢包虚拟链路(CLVL, Controlled-Loss Virtual Link)的概念。通过对流束（Bundle）聚集速度的控制，丢包率就能保持在一个很小的范围内。因此，无论网络环境如何变化 OverQoS 都可以获得一个下限服务质量服务，从而在统计意义上实现对带宽和丢包率指标的保证。

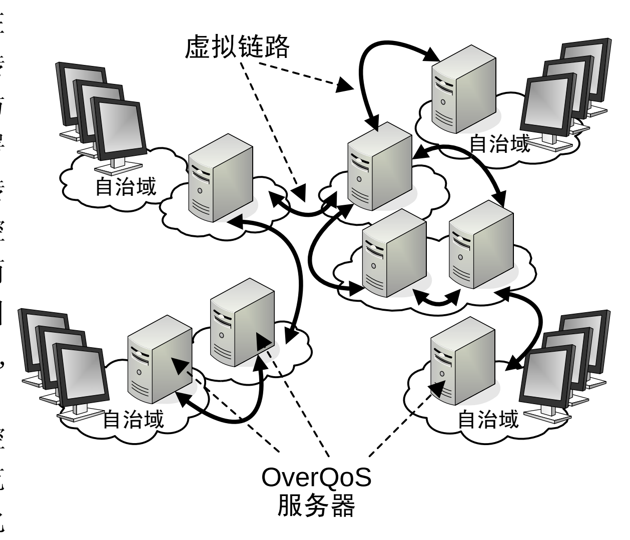


图4. OverQoS 系统架构

OverQoS 网络属于网络层覆盖网，因此存在和 Detour 一样的部署代价问题。同时，OverQoS 对流传输的优化主要是通过平滑丢包、指定数据包的优先级在统计意义上保证带宽而没有考虑其他的传输性能指标，这样对于应用的传输优化效果就有所限制。

■ SON（Service Overlay Networks）

SON[24]是一个用来解决端到端服务质量、构建和部署服务质量敏感业务的覆盖网络。它通过从各个网络域中购买含有服务质量保障的网络带宽，提供一个逻辑上的端到端的服务传输平台 SON；SON 通过与用户签订服务协议和收取费用，向用户提供增值的网络服务。SON 的结构思想也最接近于虚拟网络模型。

3 面向域间路由优化的覆盖网

尽管现有研究表明，路由覆盖网可以通过利用互联网的冗余路径，采用多种方式在一定程度上改善业务传输质量，但真正具有可部署性的针对域间路由优化的覆盖网却还面临许多现实性的问题需要解决。首先，在实现难度方面，要解决现有互联网的域间路由问题。在大量可供选择的技术方案当中，覆盖网方式需要与重建协议层或采用多穴接入(multi-homing)[4]等技术广泛比较，以确立自己的优势地位；其次，从管理角度而言，路由覆盖网作为互联网的补充，一旦大规模部署，将和互联网面临同样的扩展性、可管性问题，需要谨慎考虑其流量优化功能是否会与各互联网服务提供商对 IP 层的流量管理策略以及与其他覆盖网冲突而引起网络震荡；再次，覆盖网方案对网络流量性能的自身改善能力与开销还需要进一步量化对比；最后，覆盖网技术还存在是通过升级路由器还是部署终端服务器实现的分歧。

3.1 域间路由优化覆盖网与其余优化方案对比

针对互联网当前传输服务方面所出现的问题，其解决方案一直存在较大的分歧：一类方案认为应该完全推倒重来，如 GENI、FIRE 等，一类方案认为应该完全兼容现有网络体系结构，如覆盖网、多穴接入等；还有一类方案取两者的折中，认为可以进行一些局部架构的革新设计。

[4] 有译作“多链路”或“多归属”，即一个主机拥有一个以上的 IP 地址。

3.1.1 覆盖网 vs GENI、FIRE

GENI 是由美国国家自然科学基金会 NSF(National Science Foundation)主持推动的建设下一代试验研究网络项目[37]。与覆盖网满足兼容、过渡的要求不同，GENI 考虑的是“15 年后”的全球网络需求，以及如何不受约束从头开始建设网络(Clean-slate)。其研究方向不局限于基础设施，还包括了原理、协议、体系结构和方案设计。目前参与 GENI 的学校和研究所几乎囊括了美国所有顶尖的机构：斯坦福、麻省理工、普林斯顿、甚至国防部等等。而积极参与的公司有思科、富士通、惠普、微软研究院、NEC 等大量知名国际企业。GENI 对体系结构的核心概念要求是：可编程(Programmability)、可虚拟化和资源共享。

欧盟在网络架构革新设计方向上，也有一个类似的项目，叫做 FIRE（Future Internet Research and Experimentation）[38]。FIRE 的原则是推进基于试验的研究，把学术界前瞻研究和工业界测试实验两端相结合。同样，FIRE 也在欧洲致力创建大规模的、动态的、可持续的实验设施，这个过程中，将把目前各个小规模的试验床链接和联合起来。

但无论是 GENI 还是 FIRE，这些“从一张白纸开始”式的国家级研究努力还处在初级阶段——在 10 到 15 年之内，可能很难指望产生出有意义的成果。相比覆盖网技术，GENI 和 FIRE 都是一种全新的革命(Revolution)，但从实践出发，演化(Evolution)可能是更现实的选择。因为现在的互联网远超出学术界所有，另起炉灶的方案今天比四十年前互联网初创面临的困难要多得多。

3.1.2 覆盖网 vs 多穴接入

相对覆盖网的部署，Multi-homing 是一种更为简单的性能优化方式。大型企业或小型互联网服务提供商通常通过这种方式连接到互联网，以取得更可靠、稳定的路由性能。但使用多穴接入的方式接入互联网，必须支付给互联网服务提供商冗余链路使用费，其优化效果与开销成正比。戈登伯格（D.K. Goldenberg）等通过建立费用开销模型，结合时延等性能指标的优化，设计了一套智能多穴接入算法[39]， 通过这套算法，可以在单连接的基础上，以增加一个互联网服务提供商的代价，换取 18%左右的网络性能提升。

阿克拉（A. Akella）则对多穴接入与覆盖网在往返延迟时间(Round-trip time，RTT)、吞吐量等方面相对缺省路由的优化能力做了详细对比[40] 。其测试表明，在单个覆盖网与一个多穴接入的情况下，覆盖网具有更好的优化能力，其中时延指标平均相差 33%，吞吐量相差 15%。而在 K 个多穴接入(K>3)的情况下，相对单个覆盖网，多穴接入的网络性能略优于覆盖网技术。如果在多穴接入(K>3)的基础上再使用覆盖网技术，其性能提升非常有限(10%左右的时延改善，5%左右的吞吐量改善)。同时，互联网服务提供商要使用覆盖网技术，还需要彼此签订复杂的流量转发协议。因此，阿克拉认为，互联网服务提供商完全可以通过基于多穴接入的智能路由机制来提升网络性能，而无需专门架设覆盖网络。

但实际上，虽然多穴接入的实现简单，但多个接入会造成客户端成倍的开销，而覆盖网的规模扩大后，可以提供大量的离散路径，进一步提高优化的性价比。因此，朱永（音译，Yong Zhu）等在[41]中提出可以将两者的优点结合起来，形成一个 MON(Multihomed Overlay Network)网络，并针对 MON 的部署位置、设计方式、运营开销等做了深入的讨论。其理论分析的结果表明，采用覆盖服务提供商(Overlay Service Provider，OSP)方式运营，相对传统互联网服务提供商，不仅可以提高网络的综合性能，还可以降低运营成本，提升盈利空间。但由于 MON 设计方式是复杂的 NP 问题，要获得最佳优化结果，还需要对用户、流量、以及性能等全局信息进行收集和分析，这在一定程度上会造成其应用的限制。因此，冈田（H. Okada）等进一步研究了多个覆盖网并存情况下的合作路由方式，并提出了一套较简单的协作协议，但其效果尚需要在真实网络环境下进行进一步验证[42]。

3.1.3 覆盖网 vs 其余新结构

受覆盖网技术的主动路由特性和多穴接入的启发，研究人员提出了一些折中的网络改造方案，较为典型的代表是杨小伟（音译，X.W. Yang）等提出的 NIRA(New Internet Routing Architecture)架构[43]。

与覆盖网技术类似，NIRA 是一种面向用户的主动路由模式。通过设计一个新的互联网域间路由体系结构，NIRA 使终端用户或应用程序可以选择其数据包所经过的一系列的互联网服务提供商。通过在域的层次上赋予终端对路由的更多控制权，一方面可以迫使互联网服务提供商之间互相竞争，从而提高服务水平，另一方面也可以带来技术上的效益，改善路由可靠性。

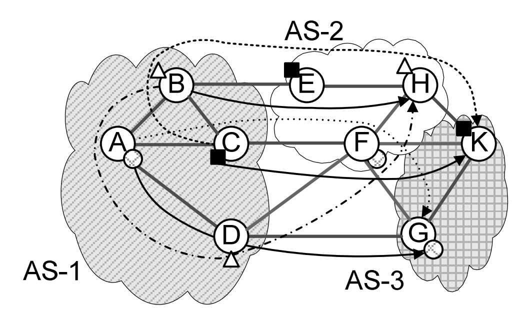
NIRA 使用 TIPP(Topology Information Propagation)进行拓扑信息的传播，实现路由发现；采用层次化编址，可实现高效的路由描述；当用户启动通信时，采用 NRLS(Name-to-Route Lookup Service)进行名字与目标路由片段的查询转换。NIRA 基于这些主要技术开发了一个可以支持用户进行路由选择的完整的路由系统。

虽然 NIRA 汲取了多穴接入和覆盖网技术在路由改善方面的研究经验[44]，避免了各自的不足，其基于层次化编址方式可大大减小路由表所记录的路由状态，解决了扩展性的问题。但 NIRA 需要在骨干网上部署，开销很大。另外它的终端用户需要进行复杂的配置，用户也难以预知服务质量，以及存在暴露互联网服务提供商路由的商业风险等。这些因素都限制了 NIRA 的应用和扩展。

3.2 跨域路由优化覆盖网的网络冲突

跨域路由优化覆盖网通常由分布在不同网络域内的预备节点构成，通过应用层路由或者虚拟路由网络进行服务质量控制。其思想就是将路由决策交由应用层，由应用层根据业务端到端时延、丢包率或吞吐量等自身需求来选取相应的路径，从而打破 IP 层对路由的垄断权。

但允许业务层与 IP 层同时控制流量路由将可能引发各个独立的流量管理与优化策略之间的矛盾与冲突[27]。如在图 5 所示的网络结构中，覆盖网 1 中的通信节点对 A-G 的备用路径 A→C→F→G 与覆盖网 2 中的通信节点对 C-K 间存在重叠部分（C→F），而 C-K 的备用路径与覆盖网 3 中节点对 B-H 间也存在重叠部分(B→E)，在 D-G 路径发生故障或拥塞时，覆盖网 1 将把其上层流量转移到备用路径(A→C→F→G)上，从而增加链路 C-F 的流量负荷，如果这种流量迁移降低了 C-K 之间的通信质量，那么覆盖网 2 又会将其流量迁移至它的备用路径(C→B→E→H→K)，同样，由于覆盖网 2 的备用路径中存在与覆盖网 3 中 B-H 通信节点对之间缺省路径重叠的链路 B-E，一旦这种迁移影响了覆盖网 3 中 B-H 间的通信，覆盖网 3 又将流量转移至其备用路径(B→A→D→F→H)。这仅仅是在 3 个覆盖网和 9 节点的简单拓扑结构下，链路性能变化所引起的流量大迁移。在覆盖网数量更多、拓扑规模更大的情况下，这种单事件变化所触发的流量波动更为复杂，甚至会引起路径反复切换导致的大规模网络震荡[28]。



覆盖网1　覆盖网2　覆盖网3
原始路径　替代路径1　替代路径2
替代路径3

	覆盖网 1（A→G）	覆盖网 2（C→K）	覆盖网 3（B→H）
原路径	A→D→G	C→F→K	B→E→H
替代路径	A→C→F→G	C→B→E→H→K	B→A→D→F→H

图5. 覆盖网下流量竞争引发的网络冲突

我们可以把这种覆盖网络中流量优化引起的网络冲突概括为两类：平行冲突与垂直冲突。平行冲突即上图中一个覆盖网流量与其余覆盖网流量或与背景流量之间的冲突；而垂直

冲突是指覆盖网流量与互联网服务提供商的流量管理策略之间的冲突。互联网服务提供商使用流量工程（traffic engineering，TE）策略应对网络性能的动态变化(如链路故障失效、BGP失效、流量拥塞等)。流量工程一般通过对流量矩阵（traffic matrix，TM)的动态估计来进行流量均衡[30]。互联网服务提供商通常对流量工程进行了两个假设：（1）流量需求不会在短时间内发生重大的改变；（2）改变域内的传输路径不会改变流量自身的需求。而域间路由覆盖网恰恰破坏了这两个假设。

3.2.1 多覆盖网之间的平行冲突

对于这种覆盖网优化所产生的平行冲突与垂直冲突问题，大量研究人员做了深入的理论分析与实验验证。在平行冲突方面，多数研究人员认为基于贪心优化策略的各个覆盖网共存竞争会降低网络的整体性能[31][32][33]。在此基础上，姜文杰（音译，Wenjie Jiang）等人使用非合作的博弈模型，在共享底层链路的多个覆盖网中对整体优化策略、基于覆盖网的优化策略与自私优化策略进一步进行了理论分析与仿真实验。其结果认为基于覆盖网的优化策略与最优解（整体优化策略）的性能较为接近，而现有覆盖网所采取的自私优化策略无论是网络整体性能还是单个覆盖网的个体业务性能均差于前述两种策略[34]。

而裘（音译，Qiu L）在采用 OSPF[5]协议的单个自治系统环境下的仿真结果却表明[29]，在平行冲突方面，覆盖网并不如前述研究结论所显示的那样使性能变得更加恶化。在类互联网(Internte-Like)环境下，彼此独立的各个覆盖网采用自私路由的优化方式后，在博弈平衡状态下，其平均网络时延与链路利用效率最终都较接近最优结果，但会导致一些最短路径频繁过载所产生的系统开销。即使增加覆盖网的数量，对端到端平均时延也只有非常轻微的影响。这表明采用现有覆盖网的自私路由方式后，其平行冲突对其性能下降的影响并不重要，裘的仿真结果还表明，在网络层路由机制设置合理时，不同覆盖网路由完全可以实现很好的共存。

这些研究结果产生分歧的根本原因在于比较对象的不同和仿真环境的差异。认为覆盖网的自私路由方式会产生整体性能恶化的研究者们，其对比算法的方案都基于全局信息共享与集中式控制管理的前提，而且基本不考虑网络层协议(如 OSPF、MPLS[6]等)对覆盖网优化算法的影响。而裘等人在衡量多个覆盖网优化中平行冲突对性能的影响时，采用的是接近真实互联网的仿真环境，在时延与网络吞吐量等指标性能上，比较的对象是基于信息共享的优化控制与互联网的缺省被动路由两种方式。

显然，要在当前互联网环境下实现大规模的全局细节信息共享或集中式路由，其可行性将受到极大制约。相对而言，实现一套针对具体业务路由需求的独立的覆盖网系统显得更有实际意义。因此，克拉拉普拉（Ram Keralapura）等针对覆盖网共存现象，进一步就其资源同步竞争问题进行了深入的研究[28]。其结果认为，即使覆盖网之间采用不同的路径性能探测方式，仍然可能导致发生对最优链路的同步竞争。这种竞争不仅影响各个覆盖网的优化性能，也会影响网络中的非覆盖网流量。而通过在探测算法中加入随机参数，并采用指数退避式的资源竞争算法可以较有效地抑制覆盖网同步概率和减少同步震荡。

3.2.2 覆盖网与互联网服务提供商网络之间的垂直冲突

在覆盖网应用是否会与互联网服务提供商网络的流量工程产生冲突而降低网络性能这个问题上，研究人员的结论比较一致。裘的研究表明，引入动态流量矩阵 Tt(s,d)后，由于覆盖网对自身流量的主动性调整，导致流量矩阵变换频繁，加大了流量工程策略进行底层优化的难度，其各自优化发生冲突的结果反而降低了网络的整体性能。同时，裘还发现，采用 MPLS 协议进行底层流量优化时，相对 OSPF 而言与覆盖网具有较好的共存能力。其原因是

---

[5] Open Shortest Path First，是一个内部网关协议(Interior Gateway Protocol,简称 IGP)，用于在单一自治系统内决策路由。

[6] Multiprotocol Label Switching，多协议标志交换协议

MPLS 对路由流量有更丰富的调节手段。

刘（音译，Liu Y）等构建了一个与底层网络完全隔离、采用独立优化策略的覆盖网系统仿真环境[35]，并假定覆盖网与互联网服务提供商的流量工程进行同频纳什博弈[7]。在此基础上发现，如果覆盖网系统频繁优化其性能，反而会使得其开销大大增加，最终导致优化性能下降。这是因为在覆盖网选择路径后，动态流量矩阵将调整自身路由，以使整体网络开销最小化，而更新的流量工程策略将增加覆盖网的开销，这种交互博弈将反复进行，直到达到纳什平衡。Liu Y 的仿真实验同样表明，相对 OSPF，使用 MPLS 协议的流量工程，可取得较好的优化效果。同时，在覆盖网流量占据主导地位时，其路由决策对流量工程的开销影响非常明显。

尽管这些研究结果表明，互联网服务提供商采用基于 MPLS 的流量工程策略后，理论上有可能避免覆盖网进行流量优化时产生的垂直冲突，但在实际应用上仍然存在许多挑战性问题，如动态的全局流量矩阵信息获取、大规模线性编程实现等。因此，对于互联网服务提供商来说，其态度显然倾向于限制覆盖网应用的发展，以降低网络开销，阻止整体性能恶化。

但王慧（音译，J. H. Wang）等人从商业模型博弈的观点出发[36]，通过对域内域间的流量进行建模推导，发现在纳什均衡下大型互联网服务提供商将陷入两难的尴尬境地：当流量模型从传统的 Web 服务方式迁移到覆盖网方式时，不仅导致网络流量的不稳定，小型互联网服务提供商还将享受到搭便车(free-ride)的好处。而简单采取缩减或限制覆盖网端用户流量又将使互联网服务提供商业务受损，影响接入定价权，导致客户的流失；如果缩减私有链路的使用，又会影响其余业务与网络的扩张，不利于自身与互联网的发展。

总之，覆盖网已成为难以逆转的技术潮流，它的出现使得基于 BPG 协议的松散互联网耦合性加强，不同自治系统之间将会相互影响，流量矩阵变得更为动态复杂，传统的流量均衡策略也会被打破。互联网服务提供商必须充分认识覆盖网络，在此基础上采取相应的对策，进行特性规律统计，并提供良好的交互性与信息共享机制。覆盖网也应当尽可能了解互联网服务提供商的路由策略与底层信息，并采取相应的机制如随机路径探测等方式，以避开同步震荡，避免对其他覆盖网流量与网络背景流量(非覆盖网流量)造成的不利影响。

3.3 域间路由优化覆盖网性能分析

3.3.1 域间路由优化覆盖网技术的现实基础

覆盖网进行网络优化的重要手段和策略就是为应用找到更多的可靠备用路径，当必要时用户可以把应用传输的路径切换到这些备用路径，以保证传输服务质量。这个方法的前提是要求底层网络在通信节点对之间拥有多条冗余路径，特别是多条独立的物理路径。而当前的网络确实能够提供这样的条件。

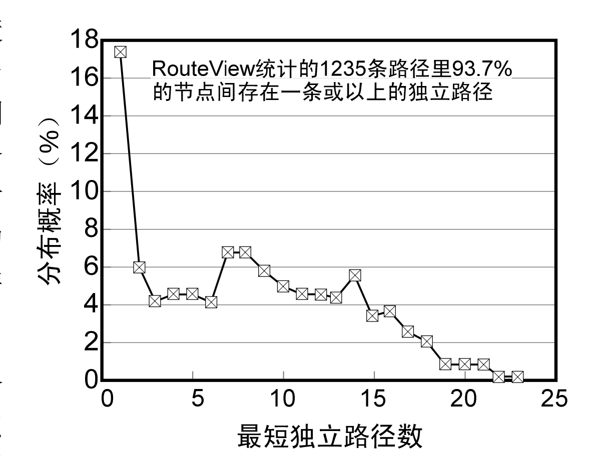


图6. 独立最短路径分布

图 6 显示的是在 RouteView 上的 42 个节点之间的 1235 条边上所测得的所有节点之间独立最短路径的一个累积分布[45]。从图中可以看出，有 17.4%的最短路径拥有 1 条独立最短路径，而 6.0%的最短路径拥有 2 条独立最短路径等等。最终统计得到有 93.7%的最短路径拥有至少 1 条独立的最短路径。这充分说明了当前的网络的路径冗余度非常高。而

[7] 非合作博弈

由于当前互联网采用的路由协议，包括域间 BGP 和域内 OSPF 等自身的局限性，这些大量存在的隐藏路径通常不能被路由协议发现和利用。比如，许多互联网服务提供商不接受少于 8192 个连续地址块发起的 BGP 通告。这样即使这些网络与其他网络之间有多个连接，但由于在 BGP 路由表中不能找到，就成为了隐藏路径而无法使用。此外，随着网络的发展，很多网络之间产生一些私有链路。这些私有链路本来也可以被用来优化传输，但是不能被当前的网络协议所找到。而覆盖网通过节点之间的探测和交互，主动地选取多条备用路径的方式，则可以发现并利用这些冗余路径。

3.3.2 域间路由优化覆盖网技术的改善能力分析

覆盖网技术相对传统的互联网路由在网络性能方面的改善程度直接关系到其研究价值与应用前景。因此，有关覆盖网实用性方面的争论一直未曾停止过。如前文所述，近 10 年来，有关覆盖网的研究大致可分为中继型覆盖网与虚拟网络型覆盖网。后者往往与优化具体业务的相应指标有关，难以进行定量评测分析。因此，大部分仿真与测试方案都集中在中继型覆盖网路由方面。

测试地点与对象	测试网规模	时延改善	丢包率改善	吞吐量改善	系统开销
a. 美国 弹性覆盖网网络 vs BGP 路由	13 个节点，包括 2 个欧洲节点	11%以上的抽样分组可改善 40ms 以上的时延	网络良好时：5%以上； 丢包率 10%时：15%以上；	10%以上	$O(N^2)$
b. 美国 覆盖网 vs 多穴接入 vs BGP 路由	68 个节点，分布在美国 17 个城市	vs 多穴接入: 5%-15% vs BPG: 33%平均时延		平均 17%	$O(N^2)$
c. 日本 覆盖网路由 vs BGP 路由	基于 plantlab 的 s3 的 588 个节点	100ms 以内的时延数比例提升 30%以上(50%→80%)		100Mb 连接数提升 10% (80%→90%)	$O(N^2)$
d. 日本 基于 Ping 探测的覆盖网路由 vs BGP 路由	18 个节点，包括东京、大阪等 4 个区域	100ms 以内的时延数比例提升 15%以上(80%→95%)			$O(N^2)$
e. 美国 Plantlab 中 Indirect 路由 vs BGP 路由	43 个节点，包括 22 个国外节点			平均 33%~49%	$O(N^2)$
f. 美国 带宽优化覆盖网路由 vs BGP 路由	基于 plantlab 的 s3 的 174 个节点	3%左右的节点对可提供低时延、高带宽的转发		平均 20%以上（40ms 以内）	$O(N^2)$

表1. 中继型覆盖网优化性能统计表

早在 2001 年，安德森（D. Andersen）等就提出了后来成为中继型覆盖网路由典型代表的弹性覆盖网。其测试结果表明，覆盖网路由的优化能力与网络环境相关(如表 1 中 a 所示)，阿克拉、奥帕斯（J. M. Opos）、李（译音，Sung-Ju Lee）在更大范围网络环境下[40][46][49]，对覆盖网与 BGP 路由性能进行了对比测试，其优化效果要比弹性覆盖网中的测试结果明显(如表 1 中 b、e、f 所示)。平冈（Hiraoka）、内田（Uchida）[47][48]在日本分别利用 Plantlab 平台和布点测量的不同方案，测试了覆盖网对网络时延与吞吐量的改善程度(如表 1 中 c、d 所示)。在所有测试方案中，覆盖网的优化方式均采用了中继转发技术，通过遍历所有转发节点对，为端到端业务提供最优的传输路径，因此其系统开销均为 $O(N^2)$。

而同一种跨域优化的覆盖网方案，甚至在同一个测试平台上，对性能的改善都呈现非常大的差异。其原因是阿克拉、奥帕斯的测试方案是非完全测试，即只选择了部分性能低下的网络节点作为测试点。这导致平均性能提高远高于其余完全测试方案。即使是采用完全测试方案，由于网络环境的不同，平冈、内田的测量在时延优化结果上也出现了很大的差异。覆盖网对真实节点平台通信的时延改善能力不及 Plantlab 平台中的表现。而 S.J.李的完全测试结果则表明，虽然覆盖网中存在大量的冗余路径，但能同时提供低时延、高带宽转发的节点数只占其中极小的比例，对吞吐量的显著提升，需要以牺牲时延性能为代价。

综合这些测试结果，可以发现覆盖网技术在网络环境越差的情况下，对系统性能改善越明显。这是由于在网络环境良好的时候，中继型覆盖网需要在中间节点的应用层或网络层进行一次或多次的转发，所产生的额外开销降低了优化效果。此外，由于存在大量的性能各异的冗余路径，覆盖网可以专门针对应用所需要的某个指标进行相应的优化，从而使相关性能得到幅度更大的提升。因此，在当前互联网健壮性、稳定性都有待完善的情况下，研究如何在应用层部署覆盖网技术，并解决其关键性问题，显然具有很大的实用价值。

4　高效的域间路由优化覆盖网实现

如前文所述，由于网络层覆盖网一般由第三方互联网服务提供商提供，如 Detour、OverQoS 等。虽然它们具有转发性能好，稳定性高的优点，但其部署灵活性与扩展性较差的缺陷也显而易见。而且，网络层覆盖网一般只针对 IP 层域间流量进行优化，对终端应用的需求无法细分。从覆盖网的可扩展性和可部署性的要求考虑，应用层实现的覆盖网如 Spines、弹性覆盖网等更具有发展前景。终端应用可利用这一类覆盖网，找到真正适合自身需求的传输路径。如：在线游戏可选用高吞吐量与低延迟的路径；文件对等传输（p2p）可选用费用低廉的路径；而对一些要求高可靠的（Mission–critical）业务，还可以在上层自建更可靠的传输协议来提升其可靠性。

但应用层覆盖网也有其自身的弱点。如现有应用基本不考虑底层网络的结构信息，这将导致在传输优化的过程中出现路径相关而带来性能下降，同时也可能引发前文所述的网络垂直冲突等问题；其次，应用层进行转发节点选取和组网时，一般采用完全网状连接探测，它的开销为 $O(N^2)$ 级别（如弹性覆盖网），这严重影响了网络的扩展性；此外，多径路由是覆盖网实现传输优化的一种重要手段[50][51]，但在如何有效减少链路动态变化对网络传输的影响，如何抑制路由切换震荡，如何进行拥塞控制及转发节点部署等方面都存在有待深入研究解决的问题。

4.1 可扩展的覆盖网系统结构

基于Chord环的自组织方式
互联网骨干
平台节点
服务器
服务节点
覆盖网络
IAD
语音接入
转发节点
用户终端
视讯接入

图7. 可扩展的覆盖网系统

考虑到部署的成本和可行性，现在大部分实用的覆盖网络采用的都是基于应用层的中继模式。这种方式由于终端主机的不稳定性，容易造成覆盖网逻辑拓扑的动荡变化，严重影响转发性能。而单纯使用网络中固定部署的转发节点构建网络又面临系统成本与灵活性的问题。

因此，可扩展的域间路由优化覆盖网系统应当充分利用两类节点的优势，采用分层划分、混合的方式构网。具体而言，就是将预部署的服务节点主要用于转发网络的分层、管理与维护，并承担部分业务转发功能。而由经过筛选后的终端主机承担主要的业务路由服务（如图7所示），它们与普通用户节点构成一个三层结构。

这个三层架构要克服各类覆盖网稳定性、部署成本、扩展性等方面的难点问题，还需要在转发节点筛选技术、激励机制、动态组网管理等方面做进一步的深入研究。

4.2 基于底层信息感知的路径选取、逻辑组网与探测

限制现有应用层覆盖网性能与规模的问题主要还在于垂直网络冲突的存在。大多数的覆盖网基本不考虑网络底层结构与流量等信息，它们独立维护各自的逻辑网络，通过洪泛方式监测所有可能的备选路径质量来寻找有效的替换路径。如弹性覆盖网就是通过固定时间发送探测包从而试图尽可能快地恢复路径故障或拥塞问题。这种小间隔、频繁的探测给网络带来了$O(N^2)$的巨大开销。因此，弹性覆盖网的节点规模数不能超过50以上。

4.2.1 基于路径独立性的转发节点选取

尽管如此，在实际网络中仍有大约 40%－50%的路径故障通过这种应用层覆盖网络不可恢复。这是由于覆盖网备用路径与默认的底层 IP 路径故障相关联。也就是说覆盖网路径与底层的 IP 路径所共享的物理链路或路由器失效后，导致覆盖网使用的替代路径也同时失效。实际上，这也是优化覆盖网网络路由所需要考虑的最根本问题。

汉（J.Han）的研究[21]表明如果随机挑选覆盖网节点而不考虑底层的结构信息，将会导致大量的覆盖网路径重叠相关。因此，尽管使用很小的间隔来探测路径性能，覆盖网络对于快速应对故障和拥塞的能力还是有限的，除非覆盖网路径可以保证在 IP 层的元素完全分离，不存在相关性。这就提出了如何构建覆盖网路径使得他们之间或者与默认路径在底层 IP 网络之间相关性最小化的问题。Cha 提出一种通过增量部署中继节点来构建覆盖网路径的算法[54]，其目标是减少共同存在于缺省路径和覆盖网路径上的链路数目。该方法假设域内的网络拓扑已知，并在此基础上分析优化域内中继节点的放置。类似的，通过获取拓扑信息来

选取合适的中继节点以构建覆盖网路径的目标也在于减少默认路径与覆盖网路径之间重叠的网络元素数。

4.2.2 探测方式与基于 n-Landmark 的逻辑组网

针对频繁的路径探测所引起的扩展性问题，瑞瓦斯卡尔（Rewaskar） 研究了覆盖网网络的性能收益与开销间的权衡后，认为采用减少逻辑连接数、延长信息交换间隔等方式可以提升覆盖网的性能/开销比[53]。长谷川（Hasegawa）将覆盖网依据是否存在底层路径的重合分为高密度和低密度两类，并提出了在高密度覆盖网网络中，通过使用多段路径的叠加来减少探测开销的方法[56]。而程（音译，Cheng）[55]提出的 PPRR（Path Probing Relay Routing，路径探测中继路由）的探测方案最具实用性，相比弹性覆盖网等传统覆盖网采用完全网状连接探测方式后引起的$O\left(N^2\right)$开销，PPRR 的网络开销与规模无关。其原理是对每一个覆盖网节点，维护一个由其中继节点组成的“探测集（Probe set）”（如 A、B、C、D、E、F），在每轮探测中，都将当前性能最好的几个（如 A、B 和 C）节点放入“最佳集（Top set）”中，并将处于“探测集”内和“最佳集”外的其余覆盖网中继节点（D、E 和 F）全部替换为“探测集”外的覆盖网节点，再准备下一轮的路径探测。

通过实验证明，该方案可在减少探测开销的情况下，取得近似于弹性覆盖网的性能。但该方案仍存在明显的缺陷：第一、虽然每次探测的节点数量有所下降，但每个节点仍要维护一张包含网络全部节点的列表，这在大规模网络内，会影响可扩展性；第二、由于缩减了探测范围，部分路径故障信息无法及时获得；第三、所探测的覆盖网节点选取没有考虑路径相关性问题，会导致路径替换的低效。

针对这些问题，为了对备用覆盖网路径实现高效探测，提升覆盖网络的扩展性，我们认为可采用结合地标（Landmark）技术的按需探测(on-demand)方式。系统可设立 n 个地标节点。当新节点加入时，与这 n 个地标节点进行网络测距，同时在测距空间中选取网络距离较小的点做为转发节点，具体如只考虑延迟指标，则可以通过计算：

$$D_{Av}=\sqrt{\left(D_{A_{L1}}-D_{B_{L1}}\right)^2+\left(D_{A_{L2}}-D_{B_{L2}}\right)^2+\cdots+\left(D_{A_{Ln}}-D_{B_{Ln}}\right)^2}$$

来选取延迟相似度 D_{Av} 较小的节点。其中 $D_{A_{Li}}$、 $D_{B_{Li}}$ 分别是用户节点与转发节点到第 i 个地标节点的网络距离。同时，再以 n 个地标点作为目标节点，来衡量转发节点 B 的路径重叠性。利用这两个指标为用户建立一个逻辑转发网络（LFN， Logical Forward Network），最后，在建立通信时，才由端节点对逻辑转发网络中的中继进行最后的性能探测，依据即时性能来进行路径切换或多径路由。

这种“按需”的路径探测方式，完全避免了不必要的路径探测开销，通过对地标节点的探测与路径重叠性指标筛选，完成逻辑转发网络的静态构建。而由于逻辑转发网络在构建时已经满足了许多覆盖网转发优化的需求，用户节点只需要根据当时通信的需要做最后筛选，因此有效缩减了路径探测的范围，而且在默认路径出现故障或性能下降时，可大大提高备用路径的性能改善效果。

4.3 跨域覆盖网中的多径路由技术

现有 IP 路由协议在源目的之间是采用单路径进行数据转发。单路径路由的缺陷是数据吞吐量会受到现有路由策略的限制。这意味着即使存在可代替的高带宽路径，数据也可能因策略限制，在低带宽路径上传输。而且，单一路径路由也不适合于无线自组网(ad-hoc)网络。在这种环境中由于较高的路由失效率，会造成数据传输的脆弱性。

一个新颖的解决方案就是发展多径覆盖网路由技术。覆盖网中的多径路由算法可以把源、目的端数据进行分离、重组。该算法可以提高网络吞吐率，缩短网络时延，均衡负载流量，并且提高链路的健壮性。如图 8 所示，在源节点到目的节点之间，如果互联网服务提供商域间使用传统的单径 BGP 路由，即使选择最优路径，最高传输带宽也只能达到 20Mb。而使用多径路由技术，通过三条路由同时推送数据，则最高传输带宽可以达到 20+8+10=38M。

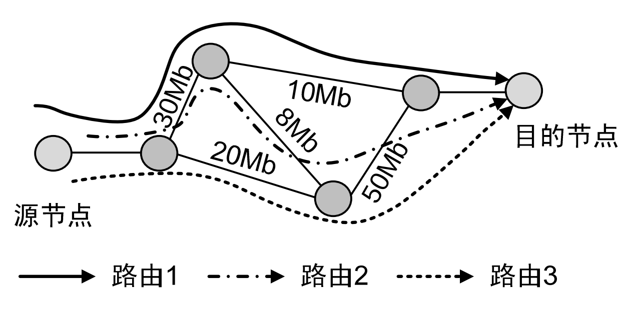


图8. 多径路由技术示意图

多径路由还可以进一步结合分层编码机制，通过附加部分冗余信息，为实时业务提供更高可靠性的传输保证[50]。但跨域覆盖网中的多径路由也存在一些问题需要解决，其中最主要的有：传输特定数据需要建立的路径数计算；给定一个拓扑，如何选取合适的路径以提供相应的服务质量保障和平衡跨域的负载流量；为了保证 TCP 的稳定性，如何设计高效的多径路由协议；给定一个拓扑结构与多径路由算法，如何设计稳定的 TCP 拥塞控制机制以提升网络容量等。

4.4 故障恢复震荡抑制

设计高效可扩展的覆盖网还有一个容易被忽略的关键技术问题，就是覆盖网自身的故障恢复机制所引起的系统开销。为了在应用层提供可靠的服务，现有网络的每个层面均有其错误检测与恢复机制。当故障或失效发生时，多个协议层都将检测到它，并采用其自身机制进行恢复。一般地，低层机制反应快，而高层机制提供更好的恢复服务，以满足不同用户及应用的需求。在覆盖网中，通过随路进行跟踪探测，来提供比 BGP 路由更为快速的故障恢复。其探测频率成为系统的一个重要参数：较低的探测频率意味着较低的系统开销，但需要更长时间检测到故障的失效。高频率的探测虽然引发更高的开销，但可以使覆盖网具有更快的故障响应速度。

当覆盖网的故障恢复机制与底层故障恢复机制存在潜在的冲突时，如：流量拥塞与传输网故障产生了许多不必要的故障事件后，这种频繁探测所引起的路径切换将导致网络的不稳定，严重影响网络性能。

因此，对于路径探测周期与路径切换方式，需要实施一定的控制策略，限制路径切换振荡。如可以考虑引入迟滞双门限的指数衰减算法或随机退避机制，使得覆盖网既可避免网络性能不稳定的情况下产生的路由震荡，同时又兼顾故障的快速恢复性能。

5　结论与展望

由于互联网的分布式管理导致了其整体传输质量无法保证的缺陷。在革新性技术与传统改进方案均遭遇到障碍后，域间路由优化覆盖网成为解决这一问题的热点技术。本文回顾了互联网路由技术现状，对现有覆盖网类型进行了细化归纳，详细分析了各类技术的优点与缺陷，在此基础上，深入探讨了应用域间路由优化覆盖网技术所面临的网络冲突、流量震荡等各种问题。同时，通过对大量覆盖网研究相关文献的仿真与实验数据的分析，我们对其路由性能改善做了全面的量化综合归纳。结果表明，在现有互联网的网络环境下，覆盖网对路由性能的改善具有较明显的效果。并且，即使在多个覆盖网共存时，也可以获得好于缺省路由的传输性能。因此，互联网服务提供商等互联网利益攸关者应加强对覆盖网技术演化的关注和重视，尽可能在两个不同层次的路由优化上提供良好的交互和协调技术，以减少网络垂直

冲突，保证互联网性能。最后，我们从系统结构、路径探测、震荡抑制、多径路由等几个方面总结了实现高效、可扩展的域间路由优化覆盖网的关键技术问题，并给出了相关建议。

参考文献:

[1] S. Blake, D. Black, and et al. An architecture for Differentiated Services. IETF RFC 2475, 1998.

[2] R. Braden, D. Clark, and S. Shenker. Integrated Services in the internet architecture: an overview.IETF RFC 1633, 1994.

[3] V. Paxson. End-to-End Routing Behavior in the Internet. Proc. ACM SIGCOMM,1996.

[4] C. Labovitz, R. Malan and F. Jahanian. Internet routing instability. *IEEE/ACM Trans Networking*, vol. 6, no. 5, pp. 515-558.1998

[5] W. Li. Inter-domain routing: Problems and solutions. Technical Report TR-128. *Department of Computer Science, State University of New York*. Feb, 2003

[6] S. Savage et al., "The end-to-end effects of Internet path selection", *Proc. ACM SIGCOMM*, pp. 289-99, 1999

[7] C. Labovitz, A. Ahuja, A. Bose and F. Jahanian. Delayed Internet routing convergence. *Proc. ACM SIGCOMM*. 2000

[8] A.Bremler-Barr et al., "Improved BGP Convergence via Ghost Flushing", *Proc.IEEE INFOCOM*, 2003

[9] D. Pei et al., "BGP-RCN: improving BGP convergence through root cause notification," Comput. Netw. ISDN Syst., 48(2):175-94, 2004

[10] W. Xu and J. Rexford, "MIRO: multi-path interdomain routing", *Proc. ACM SIGCOMM*, pp. 171-82, 2006

[11] Y. Chen, R. Katz and J. Kubiatowicz. Dynamic Replica Placement for Scalable Content Delivery. *Proc. of International Workshop on P eer-to-Peer Sytems*, 2002.

[12] Y. Chu, S. G. Rao, S. Seshan, H.Zhang. Enabling conferencing applications on the internet using an overlay multicast architecture. *Proceedings of ACM SIGCOMM*, August, 2001

[13] J. Jannotti, D. K. Giford, K. L. Johnson, et al. Overcast: reliable multicasting with an overlay network. *Proceedings of 4th Symposium on operating Systems Design and Implementation(OSDI '00)*, October 2000

[14] G. Fox. Peer-To-Peer Networks. *Web Computing*, 2001.5-6.

[15] http://www.planet-lab.org

[16] B. Raman, S. Agarwal, Y. Chen et al. The SAHARA Model for Service Composition Across Multiple Providers. *Pervasive Computing*, August 2002

[17] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient overlay networks. *In Proc. ACM Symp. Operating Syst.* Principles,2001, pp. 131–145.

[18] Zhi Li, Prasant Mohapatra. QRON: QoS-Aware Routing in Overlay Networks. *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, VOL. 22, NO. 1, JANUARY 2004.

[19] Subramanian L,Stoica I, Balakrishnan H,et al. OverQoS: An overlay based architecture for enhancing internet QoS. In:Proceedings of USENIX 1st Symposium on Networked System Design and Implementation (NSDI2004).San Francisco:USENIX Press,2004.71-84.

[20] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A.Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, and J. Zahorjan. Detour: Informed Internet routing and transport. *IEEE Micro*, vol.19, no. 1, pp. 50–59, Jan. 1999.

[21] J. Han, D. Watson, F. Jahanian. Topology aware overlay networks. *IEEE INFOCOM* 2005

[22] Z. Li and P. Mohapatra. QRON: QoS-aware routing in overlay networks. *IEEE Journal on Selected*

Areas in Communications, vol. 22, no. 1, pp. 29–40, 2004.

[23] Subramanian L,Stoica I,Balakrishnan H,et al. OverQoS: An overlay based architecture for enhancing internet QoS. In:Proceedings of USENIX 1st Symposium on Networked System Design and Implementation (NSDI 2004).San Francisco:USENIX Press,2004.71-84

[24] Z. Duan, Z. L. Zhang and T. Hou. Service Overlay networks: SLAs, QoS and Bandwidth Provisioning. *Proc. of 10th IEEE International Conference on Network Protocols(ICNP2002)*, Paris, France, November 2002.

[25] D. Xu and K. Nahrstedt. Finding Service Paths in a Media Service Proxy Network. *Proc. of SPIE/ACM Multimedia Computing and Networking Conference, San Jose, CA*, January2002.

[26] D. Xu, K. Nahrstedt, and D. Wichadakul. QoS and Contention Aware Multi-Resource Reservation. Cluster Computing, the Journal of Networks, Software Tools and Applications, 4(2), Kluwer Academic Publishers, 2001.

[27] Ram Keralapura1;2, Nina Taft2, ChenNee Chuah1, Gianluca Iannaccone2，Can ISPs Take the Heat from Overlay Networks? *ACM HotNets Workshop*. November, 2004. San Diego

[28] Ram Keralapura, Chen-Nee Chuah, Race Conditions in Coexisting Overlay Networks，*IEEE/ACM TRANSACTIONS ON NETWORKING*, VOL. 16, NO. 1, FEBRUARY 2008

[29] Qiu L, Yang YR, Zhang Y, Shenker S. On selfish routing in Internet-like environments. In: Feldmann A, Zitterbart M, Crowcroft J, Wetherall D, eds. *Proc. of the ACM SIGCOMM 2003*

[30] A. Soule, A. Nucci, E. Leonardi, R. Cruz, and N. Taft. How to Identify and Estimate the Largest Traffic Matrix Elements in a Dynamic Environment. *In ACM SIGMETRICS*, June 2004.

[31] T. Roughgarden and E. Tardos. How bad is selfish routing? *Journal of ACM*, 49(2):236–259, 2002.

[32] R. FeldmannM. GairingThomas LueckingBurkhard MonienManuel Rode，Selfish Routing in Non-cooperative Networks: A Survey，MFCS 2003

[33] R. Gao, C. Dovrolis, and E. W. Zegura, “Avoiding oscillations due to intelligent route control systems,” in *Proc. IEEE INFOCOM*, 2006

[34] Wenjie Jiang,, Dah-Ming Chiu, John C.S. Lui, On the interaction of multiple overlay routing,Performance Evaluation 62 (2005) 229–246

[35] Liu Y, Zhang HG, Gong WB, Towsley DF. On the interaction between overlay routing and underlay routing. In: Makki K, Knightly E, eds. *Proc. of the IEEE INFOCOM 2005*. Piscataway: IEEE, 2005. 2543-2553

[36] Jessie Hui Wang, Dah Ming Chiu b,*, John C.S. Lui c"A game–theoretic analysis of the implications of overlay network traffic on ISP peering" Computer Networks 52 (2008) 2961–2974

[37] http://www.geni.net

[38] http://cordis.europa.eu/fp7/ict/fire/

[39] D. K. Goldenberg, L. Qiu, H. Xie, Y. R. Yang, and Y. Zhang, “Optimizing cost and performance for multihoming,” in SIGCOMM, 2004.

[40] A Akella, J Pang, B Maggs, S Seshan ，A Comparison of Overlay Routing and Multihoming Route Control，SIGCOMM’04, Aug. 30–Sept. 3, 2004, Portland, Oregon, USA.

[41] Yong Zhu, Constantine Dovrolis, and Mostafa Ammar，Combining multihoming with overlay routing,*IEEE INFOCOM 2007*

[42] Hiroki Okada, Tran Nguyen Trung, Kazuhiko Kinoshita, and Koso Murakami,A Cooperative Routing Method for Multiple Overlay Networks, IEEE2009

[43] Xiaowei Yang,David Clark, Arthur W. Berger, NIRA: a new inter-domain routing architecture，*IEEE/ACM Transactions on Networking (TON)* Volume 15, Issue 4 (August 2007) Pages: 775 – 788

[44] K. P. Gummadi, H. Madhyastha, S. D. Gribble, H. M. Levy, and D.J. Wetherall, "Improving the reliability of internet paths with one-hop source routing," in *Proc. OSDI*, 2004, pp. 183–198.

[45] Akihiro Nakao, Larry Peterson, Andy Bavier. A Routing Underlay for Overlay Networks. *SIGCOMM'03*, August, 2003

[46] J. M. Opos, S. Ramabhadran, A. Terry, J. Pasquale, A. C. Snoeren, and A. Vahdat. A performance analysis of indirect routing. *In Proceedings of the IEEE IPDPS 2007.*

[47] Yuichiro Hiraoka, Go Hasegawa, Masayuki Murata, Effectiveness of overlay routing based on delay and bandwidth information,2007 *Australasian Telecommunication Networks and Applications Conference,December 2nd* – 5th 2007, Christchurch, New Zealand

[48] Masato Uchida,Satoshi Kamei, and Ryoichi Kawahara,Performance Evaluation of QoS-Aware Routing in Overlay Network,ICOIN 2006, LNCS 3961, pp. 925–934, 2006.

[49] Sung-Ju Lee, Sujata Banerjee, Puneet Sharma, Praveen Yalagandula, and Sujoy Basu，Bandwidth-Aware Routing in Overlay Networks，*IEEE INFOCOM 2008*

[50] J.Apostolopoulos. Reliable video communication over lossy packet networks using multiple state encoding and path diversity. *Proceeding of The International Society for Optical Engineering*, January 2001, vol. 4310, pp. 392–409.

[51] Z. Cen, M. Mutka, D. Zhu and N. Xi. An Overlay Network Transport Service for Teleoperation Systems. *Technical Report MSU-CSE-05-01, Department of Computer Science and Engineering, Michigan State University*, 2005

[52] B. Chandra, M. Dahlin, L. Gao and A. Nayate. End-to-end WAN Service Availability. Proc. 3rd USITS, San Francisco, CA, 2001. pp. 97–108.

[53] S. Rewaskar and J. Kaur. Testing the Scalability of Overlay Routing Infrastructures. Proc. of the Passive and Active Measurements Workshop, April, 2004.

[54] M. Cha, S. Moon, C. Park and A. Shaikh. Placing relay nodes for intradomain path diversity. IEEE INFOCOM, 2006.

[55] C. M. Cheng, Y. S. Huan, H. T. Kung, et al. Path probing relay routing for achieving hlgh end-to-end performance. Global Telecommunication Conference (GLOBECOM’04), Dallas: IEEE Press, 2004:1359-1365.

[56] Go Hasegawa, Masayuki Murata,Scalable and density-aware measurement strategies for overlay networks,2009 Fourth International Conference on Internet Monitoring and Protection

作者简介：

李彦君: 中国科学院计算技术研究所博士

张国清: 中国科学院计算技术研究所硕士生导师，博士